



Finanțat de Uniunea Europeană. Punctele de vedere și opiniile exprimate aparțin, însă, exclusiv autorului (autorilor) și nu reflectă neapărat punctele de vedere și opiniile Uniunii Europene sau ale Agenției Executive pentru Educație și Cultură (EACEA). Nici Uniunea Europeană, nici EACEA nu pot fi considerate responsabile pentru acestea.

STEAME ACADEMY

PLAN DE ÎNVĂȚARE ȘI CREATIVITATE DE FACILITARE A PREDĂRII (PLANUL L&C) - NIVEL 1 PROFESORI ÎN PREGĂTIRE: Text Mining: sunt aceste documente la fel?

S

T

Eng

A

M

Ent



1. Prezentare generală

Titlu	Text Mining: sunt aceste documente la fel?	
Întrebare sau temă pentru provocarea curiozității	Cum pot motoarele de căutare să găsească rezultate pentru o căutare de utilizator pe baza cuvintelor cheie? Cum identifică computerele documentele text care se concentrează pe aceleași subiecte? Cum modelează algoritmi computerizați datele nestructurate pentru procesarea digitală?	
Vârste, clase,...	16 - 18 ani	Clasele 10-12
Durată, calendar, activități	132 de ore	17 activități
Alinierea la ariile curriculare	Extragerea datelor, învățarea automată, modelarea datelor nestructurate, programarea computerelor	
Colaboratori, parteneri		
Rezumat - Sintează	Elevii sunt introduși în Data Mining și Machine Learning concentrându-se pe subiectele de bază ale procesării digitale a textului. Similitudinea textului este explorată arătând fundamentele sale matematice, intersecția mulțimilor și cosinusul dintre doi vectori. Elevii lucrează în echipe pentru a implementa un instrument simplu pentru a măsura similitudinea dintre două documente text. În ultimele sesiuni elevii sunt provocați la o competiție pentru a identifica cea mai bună implementare. În timpul tuturor sesiunilor, elevii sunt introduși în metodele cheie pentru pre-procesarea textului, cum ar fi cuvintele oprite și stemming. Ultima sesiune se încheie prin împingerea studenților să discute și să identifice asemănările dintre implementarea lor și un motor de căutare și, de acolo, să proiecteze un motor de căutare folosind implementarea lor anterioară pentru similitudinea textului.	
Referințe, Mulțumiri		

2. Cadrul de competente STEAME ACADEMY*

Cooperarea profesorilor	Profesor 1 (Științe) • Învățarea automată și extragerea datelor: prezentare generală a domeniului, arhitectură generală, aplicații comune în viața de zi cu zi, probleme comune • Modelarea datelor pentru prelucrarea digitală: reprezentarea datelor structurate pentru învățarea automată și extragerea datelor, modelarea datelor nestructurate pentru învățarea automată și extragerea datelor • Extragerea textului: prezentare generală a domeniului, concepte de bază, modelare (model boolean, model TF, model TFxIDF), similitudinea documentelor text, pre-procesare, sarcini principale și aplicații Profesor 2 (Inginerie) • Programare Python pentru extragerea textului • Algebra vectorială în Excel Profesor 3 (Matematică) • Funcția cosinus, algebră vectorială • Operatori de setare, intersecție Profesorul 1 cooperează cu Profesorul 2 și Învățătorul 3 pentru: - identificarea bibliotecilor Python de utilizat pentru extragerea textului - identificarea funcțiilor Excel de utilizat pentru algebra vectorială - Creează exercițiile și provocarea Profesorul 1 cooperează cu Învățătorul 2 pentru: - adunarea corpusurilor necesare pentru activitățile practice - adnotarea corpusurile necesare pentru fiecare exercițiu Profesorul 1 cooperează cu Profesorul 3 pentru: - introducerea operatorilor de set într-un mediu de procesare a textului - Introducerea algebra vectorială într-un mediu de procesare a textului.
STEAME în viața de zi cu zi. Organizare	Ultima sesiune se încheie prin împingerea elevilor să discute și să identifice asemănările dintre implementarea lor și un motor de căutare, cum ar fi Google, și, de acolo, să proiecteze un motor de căutare folosind implementarea anterioară pentru similitudinea textului.
Formularea Planului de Acțiune	Faza pregătitoare 1. Cercetarea aplicațiilor tradiționale și de ultimă generație pentru data mining și machine learning; link către motoarele de căutare și IA generativă; se

- referă la cazurile de date structurate și nestructurate; Revizuirea principalelor provocări ale extragerii datelor și învățării automate (modelarea datelor, multidimensionalitate, supraajustare, date lipsă, volum de date, big data, explicare versus predicție, ...)
2. Adunați și adnotați corpus pentru exerciții
 3. Configurați mediul de programare Python (docker, depozit în Github pentru clonare, altele)

Structura atelierului

1. Introducere
 - 1.1. Prezentare generală a Data Mining și Machine Learning: perspectivă istorică, sarcini/probleme, aplicații, modelarea datelor nestructurate, cum ar fi imagini și text).
2. Modelarea textului
 - 2.1. Adunați corpusurile care vor fi utilizate pentru exerciții și lucrări de proiect (trebuie să aveți documente mici cu un lexicon foarte redus; includeți documente în care TF poate face diferența în comparație cu modelul boolean).
 - 2.2. Modele pentru a reprezenta text, exemple: model boolean, TF, TFxIDF, altele. Consultați Bag of Words.
 - 2.3. Setări operatorii (uniune, intersecție etc.).
 - 2.4. Implementați mai multe versiuni ale unui instrument în Excel pentru a calcula similitudinea dintre două documente text pe baza operatorilor de mulțimi și/sau algebra vectorială (produs intern, cosinus, intersecție și uniune de mulțimi, ...).
 - 2.5. Introduceți algebra vectorială: produs intern și cosinus.
 - 2.6. Arătați impactul/relevanța frecvenței termenilor în comparație cu boolean. Clarificați faptul că cosinusul funcționează bine.
 - 2.7. Contestații de similitudine bazate doar pe fondul fiduciar (termenii prezenți în toate documentele nu au putere discriminatorie); soluții alternative.
 - 2.8. Introducerea IDF și TFxIDF; discutați, proiectați și implementați în Excel o măsură de similitudine bazată pe modelul TFxIDF pentru a fi utilizată pentru documentele de jucărie cu un lexicon redus.
3. Implementarea
 - 3.1. Cercetați R, Python sau alte biblioteci pentru extragerea textului. Pregătiți un depozit și un ghid de configurare pentru ca elevii să instaleze acest mediu de dezvoltare.
 - 3.2. Implementați o funcție pentru a calcula similitudinea dintre două documente folosind R, Python sau altele. Oferiți o implementare pentru toți elevii dacă este necesar (dacă elevii nu își pot dezvolta propria implementare în timp util).
4. Explorare matematică
 - 4.1. Implicați elevii în activități practice care explorează operatorii de mulțimi și algebra vectorială pentru a calcula similitudinea dintre

documentele text dintr-un corpus.

4.2. Facilitați discuțiile despre principiile matematice din spatele procesării textului.

5. Proiecte culminante

5.1. Proiectați proiectul (cea mai bună funcție de similitudine pentru documentele text) și redactați ghidul.

5.2. Adunați și adnotați un corpus și un set de interogări pentru a evalua precizia și amintirea produsă de implementările elevilor funcțiilor de similitudine a textului. Precizia și Revocarea vor fi evaluate din interogările/documentele de testare statice (fiecare document din corpus va fi adnotat cu clasamentul de similitudine pentru fiecare dintre cazurile de testare), rezervați un set de validare (adnotat de asemenea; acesta poate fi două sau trei documente text, fiecare cu câțiva termeni ca și cum ar fi o interogare pentru un motor de căutare; pentru fiecare dintre aceste "interogări" se va asigura clasamentul documentelor în corpusul studenților, acesta va fi folosit pentru a calcula F1 în cele din urmă și pentru a anunța câștigătorul).

5.3. Furnizați informații despre tehnicile de pre-procesare (eliminarea cuvintelor de oprire, stemming etc.) și implementarea acestora folosind bibliotecile de text mining selectate pentru proiect.

6. Link către motoarele de căutare

6.1. Pregătiți scenariul de validare. Un "document" poate fi o interogare ca cele pe care le folosim în motoarele de căutare precum Google, cum ar fi "camélias porto" sau "computer vision". Afișați în realitate folosind Google. Oferiți elevilor setul de validare, adică trei documente, fiecare cu câteva cuvinte cheie, ca și cum ar fi o interogare pentru un motor de căutare; Rulați funcțiile de similitudine pentru a oferi un clasament al documentelor din corpus și pentru a oferi studenților cel mai bun clasament. Calculați câștigătorul folosind măsura F1. Explicați elevilor ce este F1, Recall și Precizie.

Evaluare și reflecție

1. Evaluați înțelegerea și aplicarea de către elevi a operatorilor și conceptelor de algebră mulțimilor și vectoriale prin evaluări bazate pe proiecte, prezentări și reflecții scrise.
2. Încurajați elevii să reflecteze asupra experiențelor lor de învățare, evidențiind relația dintre matematică și extragerea textului.

Rugați elevii să proiecteze și să prezinte o metodologie de motor de căutare și un prototip nefuncțional folosind ceea ce au învățat.

** în curs de dezvoltare elementele finale ale cadrului*

3. Obiective și metodologii

Scopurile învățării și obiectivele

1. Înțelegeți conceptele și tehnicile generice de modelare și procesare utilizate în extragerea textului

Rezultatele învățării și rezultatele așteptate	- Explorați conexiunile interdisciplinare dintre extragerea textului, motoarele de căutare și algebra vectorială - Ilustrarea asemănării dintre documentele text și alte seturi de date nestructurate ca aplicații ale algebrei vectoriale
	Rezultate ale învățării - Discutați subiecte de nivel înalt legate de domeniile de extragere a textului și motoarelor de căutare - Descrieți relația dintre algebra vectorială, teoria mulțimilor și extragerea textului - Aplicați tehnici de bază de extragere a textului pentru a aborda cazuri de utilizare simple Rezultate așteptate - Funcția de similitudine a documentului text în R, Python sau alt limbaj de programare
Cunoștințe anterioare și condiții prealabile	- Cunoștințe fundamentale de algebră vectorială - Familiarizarea cu Excel - Abilități de bază de programare software - Utilizarea competență a instrumentelor IT
Motivație, Metodologie, Strategii, Scenarii de lucru	- Alocați elevii în echipe mici (3 sau 4 elevi). - Proiectați o soluție, implementați, testați și rafinați într-un mod iterativ. Utilizați o metodologie de dezvoltare iterativă. - Evidențiați conexiunile dintre algebra vectorială, modelele de documente, cosinus și similitudine. - Explorați motoarele de căutare pentru a prezenta relațiile dintre similitudinea textului și rezultatele motoarelor de căutare. - Ghidați studenții printr-o cale evolutivă de la cele mai simple modele (booleene) la abordări mai complexe (TFxIDF), introducând provocări pas cu pas în timp ce essay8ng în Excel cu implementări de bază pentru documente mici cu una sau două propoziții scurte și câțiva termeni diferiți, dintr-un lexicon de 10 sau 20 de termeni.

4. Pregătire și mijloace

Pregătire, aranjarea spațiului, <i>sfaturi de depășire a situațiilor dificile</i>	Atelierul se va desfășura într-o sală de clasă pentru aproximativ 20 de elevi, în grupuri de 3 până la 4 elevi. În mod ideal, amenajarea clasei va fi organizată în 5 până la 7 grupuri de mese în care elevii din fiecare echipă pot sta unul față de celălalt. Camera are nevoie de un proiector și un perete pentru prezentări tuturor și o tablă albă cu pixuri pentru a discuta idei.
Resurse, unelte,	Un depozit în GDrive, Teams, Github sau orice alt furnizor va fi pregătit în

materiale, atașamente, echipamente	prealabil cu tot mediul de programare (R, Python, ...) și corpusurile necesare pentru sesiunile practice.
	Trebuie furnizat un document care să ghideze studenții pe tot parcursul cursului/atelierului, explicând detaliile, rezultatele așteptate, evaluarea și rezultatele învățării pe sesiune.
Sănătate și siguranță	

5. Implementare

Activități educative, proceduri, reflecție	Structura atelierului 1. Introducere [1 activitate, 16 ore] 1.1. Prezentare generală a Data Mining și Machine Learning: perspectivă istorică, sarcini/probleme, aplicații, modelarea datelor nestructurate, cum ar fi imagini și text). [16 ore] 2. Modelare text [8 activități, 52 ore] 2.1. Adunați corpusurile care vor fi utilizate pentru exerciții și lucrări de proiect (trebuie să aveți documente mici cu un lexicon foarte redus; includeți documente în care TF poate face diferența în comparație cu modelul boolean). [16 ore] 2.2. Modele pentru a reprezenta text, exemple: model boolean, TF, TFxIDF, altele. Consultați Bag of Words. [8 ore] 2.3. Descrieți operatorii de mulțime relevanți (uniune, intersecție etc.). [4 ore] 2.4. Implementați mai multe versiuni ale unui instrument în Excel pentru a calcula similitudinea dintre două documente text pe baza operatorilor de mulțimi și/sau algebră vectorială (produs intern, cosinus, intersecție și uniune de mulțimi, ...). [4 ore] 2.5. Introduceți algebra vectorială: produs intern și cosinus. [4 ore] 2.6. Arătați impactul/relevanța frecvenței termenilor în comparație cu boolean. Demonstrați cu exemple. Clarificați faptul că cosinusul funcționează bine. [4 ore] 2.7. Descrieți provocările similarității atunci când se bazează numai pe TF (termenii care sunt prezenți în toate documentele nu au putere discriminatorie); soluții alternative. Demonstrați cu exemple. [4 ore] 2.8. Introducerea IDF și TFxIDF; discutarea, proiectarea și implementarea în Excel a unei măsuri de similitudine bazată pe modelul TFxIDF pentru a fi utilizată pentru documentele de jucărie cu lexicon redus. [8 ore] 3. Implementare [2 activități, 32 ore] 3.1. Cercetați R, Python sau alte biblioteci pentru extragerea textului. Pregătiți un depozit și un ghid de configurare pentru ca elevii să instaleze acest mediu de dezvoltare. [16 ore] 3.2. Implementați o funcție pentru a calcula similitudinea dintre două documente folosind R, Python sau altele. Oferiți o implementare pentru toți elevii dacă este necesar (dacă elevii nu își pot dezvolta propria
--	---

implementare în timp util). [16 ore]

4. Explorare matematică [2 activități, 4 ore]
 - 4.1. Implicați elevii în activități practice care explorează operatorii de mulțimi și algebra vectorială pentru a calcula similitudinea dintre documentele text dintr-un corpus. [2 ore]
 - 4.2. Facilitați discuțiile despre principiile matematice din spatele procesării textului. [2 ore]
5. Proiecte culminante [3 activități, 20 ore]
 - 5.1. Proiectați proiectul (cea mai bună funcție de similitudine pentru documentele text) și redactați ghidul. [8 ore]
 - 5.2. Adunați și adnotați un corpus și un set de interogări pentru a evalua precizia și amintirea produsă de implementările elevilor funcțiilor de similitudine a textului. Precizia și Revocarea vor fi evaluate din interogările/documentele de testare statice (fiecare document din corpus va fi adnotat cu clasamentul de similitudine pentru fiecare dintre cazurile de testare), rezervați un set de validare (adnotat de asemenea; acesta poate fi două sau trei documente text, fiecare cu câțiva termeni ca și cum ar fi o interogare pentru un motor de căutare; pentru fiecare dintre aceste "interogări" se va asigura clasamentul documentelor în corpusul studenților, acesta va fi folosit pentru a calcula F1 în cele din urmă și pentru a anunța câștigătorul). [8 ore]
 - 5.3. Furnizați informații despre tehnicile de pre-procesare (eliminarea cuvintelor de oprire, stemming etc.) și implementarea acestora folosind bibliotecile de text mining selectate pentru proiect. [4 ore]
6. Link către motoarele de căutare [1 activitate, 8 ore]
 - 6.1. Pregătiți scenariul de validare. Un "document" poate fi o interogare ca cele pe care le folosim în motoarele de căutare precum Google, cum ar fi "camélias porto" sau "computer vision". Afișați în realitate folosind Google. Oferiți elevilor setul de validare, adică trei documente, fiecare cu câteva cuvinte cheie, ca și cum ar fi o interogare pentru un motor de căutare; Rulați funcțiile de similitudine pentru a oferi un clasament al documentelor din corpus și pentru a oferi studenților cel mai bun clasament. Calculați câștigătorul folosind măsura F1. Explicați elevilor ce este F1, Recall și Precizie. [8 ore]

Evaluare

Evaluare și reflecție

1. Evaluați înțelegerea și aplicarea de către elevi a operatorilor și conceptelor de algebră mulțimilor și vectoriale prin evaluări bazate pe proiecte, prezentări și reflecții scrise.
2. Încurajați elevii să reflecteze asupra experiențelor lor de învățare, evidențiind relația dintre matematică și extragerea textului.
3. Rugați elevii să proiecteze și să prezinte o metodologie de motor de căutare și un prototip nefuncțional folosind ceea ce au învățat.

Prezentare - Raportare –
Împărtășirea
rezultatelor

Extensii - Alte informații

1. O funcție care calculează similitudinea dintre două documente text.

O prezentare în PowerPoint care descrie o metodologie și un prototip nefuncțional al unui nou motor de căutare propus de studenți folosind ceea ce am dezvoltat în atelier (funcția de similitudine a documentelor).

Modelul STEAME ACADEMY/Ghidul pentru abordarea învățării și creativității
Elaborarea planului de acțiune

Etapele principale ale demersului de învățare STEAME:

ETAPA I: Pregătirea de către unul sau mai mulți profesori

1. Formularea ideilor inițiale cu privire la sectoarele/zonile tematice care urmează să fie abordată
2. Implicarea mai multor actori: / angajatori / mediul de afaceri / părinți / societate / mediu înconjurător / chestiuni de etică
3. Orientarea către grupa de vârstă corespunzătoare elevilor. Concentrarea asupra conexiunilor cu programa școlară oficială. Stabilirea scopurilor și obiectivelor
4. Repartizarea sarcinilor părților implicate. Desemnarea coordonatorului și a locurilor de lucru etc.

ETAPA II: Formularea Planului de Acțiune (Pași 1-18)

Pregătire (de către profesori)

1. Raportare cu lumea reală, Reflecție
2. Stimulare. Motivare
3. Formularea unei probleme (eventual în etape sau faze) care derivă din cele de mai sus

Dezvoltare (de către elevi) – Îndrumare și evaluare (în 9-11, de către profesori)

4. Crearea contextului - Căutarea/Culegerea de informații
5. Simplificarea problemei. Structurarea problemei cu un număr limitat de cerințe
6. Definirea cazului. Proiectarea: identificarea materialelor pentru construcție/ dezvoltare/ creare
7. Construcție. Flux de lucru. Punerea în aplicare a proiectelor
8. Observare. Experimentare. Formularea concluziilor inițiale
9. Documentare. Căutarea domeniilor tematice (domenii AI) legate de subiectul studiat. Explicații bazate pe teorii existente și/sau rezultate empirice
10. Culegerea de rezultate/ informații pe baza etapelor nr. 7, 8, 9
11. Prima prezentare în grup realizată de către elevi

Configurarea și prezentarea rezultatelor (de către elevi) – Îndrumare și evaluare (din partea profesorilor)

12. Configurarea modelelor STEAME pentru a descrie / reprezenta / ilustra rezultatele
13. Studiarea rezultatelor obținute la punctul nr. 9 și formularea concluziilor, utilizând punctul 12
14. Aplicații în viața de zi cu zi - Sugestii pentru dezvoltarea etapei nr. 9 (antreprenoriat - Zile STEAME)

Revizuire (de către profesori)

15. Revizuirea problemei în general și revizuirea acesteia în condiții mai solicitant

Finalizarea proiectului (de către elevi) – Îndrumare și evaluare (de către profesori)

16. Repetați pașii 5 - 11 cu cerințe suplimentare sau noi așa cum au fost formulate la punctul 15
17. Investigare. Studii de caz. Extindere. Formularea unor ipoteze noi. Testarea noilor concluzii
18. Prezentarea concluziilor. Aplicarea unor tehnici de comunicare distincte.

ETAPA III: ACADEMIA STEAME. Acțiuni și cooperare în cadrul proiectelor creative pentru elevi

Titlul proiectului: _____

Scurtă descriere/prezentare schematică a mecanismelor de lucru/responsabilităților pentru desfășurarea acțiunii

Etape	Activități/Etape	Activități/Etape	Activități/Etape
	Profesor 1 (P1) Cooperare cu P2 și ghidarea elevilor	Acțiuni ale elevilor Grupa de vârstă: ____	Profesor 2 (P2) Cooperare cu P1 și ghidarea elevilor
A	Pregătirea etapelor nr. 1,2,3		Cooperare în cadrul etapei nr. 3
B	Ghidarea în cadrul etapei nr. 9	4,5,6,7,8,9,10	Sprijin și asistență în cadrul etapei nr. 9
C	Evaluare creativă	11	Evaluare creativă
D	Ghidare	12	Ghidare
E	Ghidare	13 (9+12)	Ghidare
F	Organizarea zilelor STEAME	14 Întâlnire cu reprezentanții mediului de afaceri	Organizarea zilelor STEAME
G	Pregătirea etapei 15		Cooperare la pasul 15
H	Ghidare	16 (repetarea etapelor 5-11)	Ghid de sprijin
I	Ghidare	17	Ghid de sprijin
K	Evaluare creativă	18	Evaluare creativă